UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 03-70
Statistics and Econometrics Series 17
December 2003

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

PARAMETRIC VERSUS NONPARAMETRIC TOLERANCE REGIONS IN DETECTION PROBLEMS

Amparo Baíllo and Antonio Cuevas*

Abstract

A major problem in statistical quality control is to detect a change in the underlying distribution of independent sequentially observed random vectors. The case where the pre-change distribution is Gaussian has been extensively analyzed. We are concerned here with the less usual non-normal multivariate case. The use of tolerance regions, defined in terms of density level sets, as detection tools arises as a natural choice in this general setup. The required level sets can be estimated in an obvious plug-in fashion, using either nonparametric or (when a parametric model is assumed) parametric density estimators. A result concerning the convergence rates of the error probabilities under a parametric model is obtained. Also, the performance of parametric and non-parametric methods is compared through a simulation study. Finally, a real data example is discussed. In general terms, we conclude that whereas the parametric estimates are, in theory, preferable when the corresponding model holds, the practical difficulties associated with their implementation make non-parametric methods a very reliable and flexible alternative.

Keywords: Level sets; statistical quality control; density estimates; normal mixtures; false alarm probability.

*Baíllo, Statistics and Econometrics Department, University Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe (Madrid), e-mail: abaillo@est-econ.uc3m.es; Cuevas, Mathematics Department, University Autónoma de Madrid, 28049 Madrid, e-mail: antonio.cuevas@uam.es. Research partially supported by Spanish grant BFM2001-0169.

Parametric versus nonparametric tolerance regions in detection problems

Amparo Baíllo
Universidad Carlos III de Madrid
28903 Getafe (Madrid), Spain

Antonio Cuevas
Universidad Autónoma de Madrid
28049 Madrid, Spain

Abstract

A major problem in statistical quality control is to detect a change in the underlying distribution of independent sequentially observed random vectors. The case where the pre-change distribution is Gaussian has been extensively analyzed. We are concerned here with the less usual non-normal multivariate case. The use of tolerance regions, defined in terms of density level sets, as detection tools arises as a natural choice in this setup. The required level sets can be estimated in an obvious plug-in fashion, using either nonparametric or (when a parametric model is assumed) parametric density estimators. A result concerning the convergence rates of the error probabilities under a parametric model is obtained. Also, the performance of parametric and non-parametric methods is compared through a simulation study. Finally, a real data example is discussed.

Keywords: Level sets, statistical quality control, density estimates, normal mixtures, false alarm probability

AMS Classification 2000: 62G15, 62G05, 62P30

Dr. Baíllo (abaillo@est-econ.uc3m.es) is Visiting Professor in Department of Statistics and Econometrics. Dr. Cuevas (antonio.cuevas@uam.es) is Full Professor in Department of Mathematics.
Research partially supported by Spanish grant BFM2001-0169.
Address for correspondence: Amparo Baíllo, Departamento de Estadística, Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe (Madrid), Spain.

Introduction

In the statistical practice one is often confronted with the need of choosing between parametric and nonparametric approaches. When the standard assumption of normality is fulfilled, the parametric Gaussian model cannot be beaten. However, in more complex situations, when a mixture of normals represents an appropriate model, the nonparametric alternatives (based, for example, on kernel estimators) are quite appealing. From a theoretical point of view it is true that, when the number of components in the mixture is known, the maximum likelihood (ML) method is asymptotically optimal and provides $\sqrt{n}$-convergence rates. Nevertheless, in practice the computational burden rapidly increases with the number of parameters to estimate and the EM-algorithm, used to implement the ML-method, can perform poorly (see McLachlan and Peel 2000), especially when the components are not clearly separated.

Here we are concerned with detection (quality control) methods based on the estimation of density level sets of type $\{f \geq c\}$ with a prespecified probability content, f being the underlying density which generates the data. Thus $\{f \geq c\}$ can be seen as a population tolerance region, which is naturally estimated by plugging in a density estimator $\hat{f}$. See, e.g., Fuchs and Kenett (1998) for a parametric (Gaussian) approach and Baíllo, Cuesta-Albertos and Cuevas (2001) for a nonparametric alternative. See also Polansky (2001) for a plug-in idea applied to evaluate the capability of a manufacturing process. Assuming a normal mixture model $\{f_\theta : \theta \in \Theta\}$ for f we have compared the performance of parametric $\hat{f} = f_{\hat{\theta}}$ versus non-parametric versions of $\hat{f}$. By "nonparametric" we mean an estimator which does not rely on any assumption of a parametric structure for f (see, e.g., the classical monograph by Silverman 1986). As a general conclusion we could say that the nonparametric methodology offers a simple, model-free and competitive choice in this type of problems.

Some background

Consider a process yielding independent observations $X_1, X_2, \ldots$ of a d-dimensional random variable X. For example, X may be certain quality characteristics of a manufactured item. In order to track the quality of the items over time, the process will be monitored for potential change in the distribution of X, by sequentially collecting and analyzing the data X_i. Initially, when the process is in control, the observations follow a distribution F. At some stage, the process may run out of control and the distribution of the X_i's changes to G. Our aim is to detect a real change in the distribution of the process as quickly as possible, subject to a bound $\alpha \in (0,1)$ on the rate of false alarms.

Let f and g denote the densities of F and G respectively. In a parametric framework, assuming normality for F with $d = 1$, the simplest detection method is the classical control chart due to Shewhart (1931). However this chart is designed for observations grouped in batches, which is not always possible or practical (see Wetherill and Brown 1994). In this work we are interested in detection methods for individual measurements. Nowadays some widely used quality control procedures for univariate observations are the Cusum chart, introduced by Page (1954), and a detection scheme proposed by Shiryaev (1963) and Roberts (1966). Both display optimality properties if f and g are known, and their performance is similar (see Pollak and Siegmund 1985).

In practice it is likely to encounter multivariate quality measurements ($d > 1$) instead of univariate ones ($d = 1$). For example, the quality of a fruit juice may be characterized by the concentrations of different aminoacids. In this setting several process variables have to be monitored simultaneously, which renders difficult the visualization and interpretation of any screening procedure. The most widely used scheme is still the T^2 control chart due to Hotelling (1947), but it imposes normality on F (see Fuchs and Kenett 1998 for a survey). Another usual practice is to screen each of the d characteristics separately, though this is inadequate when they are correlated. Some multivariate versions of the Cusum procedure have been proposed (Healy 1987, Crosier 1988, Hawkins and Olwell 1998) but, to our knowledge, only for the Gaussian distribution.

All these parametric schemes are very sensitive to misspecification of the prechange distribution F (see Gordon and Pollak 1995 for a review). For this reason some nonparametric detection procedures have been proposed in the literature, though mainly in the one-dimensional case, $d = 1$. Most of them construct a sequential test statistic based on the signs or on the signed ranks of the observations. An overview might be found in Csörgö and Horváth (1997). The generalization to the multivariate setting is not straightforward since there is no obvious and unique way of ranking multivariate observations. So far the proposals have been mainly based on the notion of depth, which induces a center-outward ordering of the sample points (see, for example, Liu and Singh 1993). Usually these approaches rely on symmetry assumptions on the density f.

Another totally nonparametric and multivariate detection procedure was suggested by Devroye and Wise (1980). Here the support $S \subseteq \mathbb{R}^d$ of F is approximated by the set $S_n(\epsilon_n) = \bigcup_{i=1}^n B(X_i, \epsilon_n)$, where ϵ_n is a smoothing parameter which converges to zero slowly enough, as $n \to \infty$. It is decided that the distribution of the process has changed at stage $n+1$ if $X_{n+1} \notin S_n(\epsilon_n)$. Baíllo, Cuevas and Justel (2000) have implemented this set-based scheme with a data-driven smoothing parameter ϵ_n chosen to control the rate of false alarms. The resulting set estimator is an asymptotic tolerance region in the sense described in the next subsection.

Our approach: asymptotic tolerance regions

The classical concept of (population) tolerance region of a distribution F as a set with a prespecified probability content $1-\alpha$ has been widely used in several fields of statistics. The key idea is that any observation supposedly coming from F should be be "most likely" within the tolerance limits of this region (see Guttman 1970). This has given rise to detection procedures in medical statistics, reliability, chemistry and certainly quality control (see Aitchison and Dunsmore 1975). As Bucchianico, Einmahl and Mushkudiani (2001) point out *"(...) in statistical process control a multivariate approach to capability studies (which, if properly conducted, should be based on tolerance regions) is highly desirable (...)"*.

Fuchs and Kenett (1987, 1988) suggested a multivariate quality control chart constructed with tolerance regions under the assumption of normality. Bucchianico, Einmahl and Mushkudiani (2001) develop a nonparametric procedure for constructing multivariate tolerance regions. Their approach, though, relies crucially on the choice of an indexing class of sets to which the tolerance region belongs (ellipsoids, hyperrectangles or convex sets, for instance). However, we are interested in populations such as mixtures of normal distributions, where it may not be that clear which class of sets to use, especially from a computational and practical point of view. Hence, our procedure will be closer to that of Chatterjee and Patra (1980). We consider tolerance regions of the type $\{f_n \geq c_n\}$, where f_n is an estimator of the original density f based on the observations $X_1, \ldots, X_n$ and c_n is such that

$$\int_{\{f_n \geq c_n\}} f_n = 1 - \alpha \quad \text{a.s.} \tag{1}$$

We will decide that the observation X_{n+1} does not follow the distribution F, that is, that $n+1$ is a change-point in the distribution of the process, if $X_{n+1} \notin \{f_n \geq c_n\}$.

Observe that, if f_n is an L^1-consistent estimator of f, that is, if $E(\int |f_n - f|) \to 0$ as $n \to \infty$ (see, e.g., Devroye and Györfi 1985), then $\{f_n \geq c_n\}$ is an asymptotic $1 - \alpha$ mean coverage tolerance region in the sense that

$$\liminf_{n\to\infty} E\left(\int_{\{f_n \geq c_n\}} f\right) \geq 1 - \alpha$$

(see Mushkudiani 2000). Chatterjee and Patra (1980) studied the behaviour of level sets $\{f_n \geq c\}$ of nonparametric density estimators f_n as tolerance regions with a guaranteed coverage.

Tolerance regions under a general parametric model

An asymptotic result

One of the advantages of the procedure introduced in the previous section is that it is applicable in both the parametric and nonparametric frameworks, depending on the type of estimator (parametric or not) used for f. In this work we are interested in checking and comparing the detection performance provided by tolerance regions when the original density f belongs to a rather general parametric family (mixture of normals) and f_n is either an estimator from that family or a nonparametric (kernel) one. Finite mixture distributions are a flexible and widely used model in a large number of fields (for a thorough survey see McLachlan and Peel 2000). They have been applied to account for the presence of contaminants among the data or to model asymmetrical distributions and in general they provide a semiparametric approximation to unknown density shapes. In the context of quality control there are also situations where a symmetric unimodal density (like the Gaussian) is not an adequate model for the underlying population. For instance, a manufacturing line frequently combines the effects of different machines, lots of material or operators and traditional control charts give troubles in these cases (DeVor, Chang and Sutherland 1992). A possible solution is to keep several separate charts, but sometimes this will not be feasible.

In this section we will first study the asymptotic behaviour of tolerance regions of the type $\{f_n \geq c_n\}$ when f_n is a parametric estimator of f and $f = f_\theta$ belongs to a parametric family. We will see that indeed this behaviour is determined by the convergence rates of the parameter estimators $\hat{\theta}$. Nevertheless the simulation results in the following subsection show that in practice the performance of nonparametric estimators is very competitive.

Let X be a random vector with density f_θ where $\theta = (\theta_1, \ldots, \theta_k)$ is an unknown parameter taking values in an open set $\Theta \subseteq \mathbb{R}^k$ $(k < \infty)$. Assume that $f_\theta(\cdot)$ is continuous on its support for any value of θ. Given $\alpha \in (0,1)$ we consider the level c such that

$$\int_{\{f_\theta \geq c\}} f_\theta = 1 - \alpha. \tag{2}$$

Let $X_1, \ldots, X_n$ be a random sample of independent observations from X and $\hat{\theta} \in \Theta$ denote some estimator of θ based on this sample (for example, the ML one). Consider the level c_n chosen as in (1). The probability of raising a false alarm is

$$P_n = \int_{\{f_{\hat{\theta}} < c_n\}} f_\theta. \tag{3}$$

A measure of the precision of the tolerance region $\{f_{\hat{\theta}} \geq c_n\}$ is given by the pseudometric

$$d_\theta(\{f_{\hat{\theta}} \geq c_n\}, \{f_\theta \geq c\}) := \int_{\{f_{\hat{\theta}} \geq c_n\} \Delta \{f_\theta \geq c\}} f_\theta, \tag{4}$$

where Δ denotes the usual symmetric difference between sets. Typically, the estimation of a level set requires to exclude the extreme case in which there is a “plateau” in $\{f = c\}$. This can be formally stated in several slightly different ways; see e.g. Tsybakov (1997), Baíllo, Cuesta-Albertos and Cuevas (2001), Baíllo (2003). We will need here a reinforced version of this assumption oriented to rule out also the existence of a “quasi-plateau” in $\{f = c\}$. In fact, we will impose that the probability induced by f exhibits at least a variation of polynomial type near $\{f = c\}$. More precisely, this is stated as follows.

$(f1)$ *There exists a constant $\gamma = \gamma(\theta) > 0$ such that $\int_{\{c \leq f_\theta \leq c+\epsilon\}} f_\theta$ and $\int_{\{c-\epsilon \leq f_\theta \leq c\}} f_\theta$ are of exact order ϵ^γ.*

In the following result we state general conditions under which the convergence of the false alarm probability (3) to α and of the pseudometric (4) to zero inherit the convergence rates in the estimation of the parameter θ.

Theorem. *Assume that* Θ *is open and convex and that, for every fixed* $x \in \mathbb{R}^d$, $f_\theta(x)$ *is differentiable with respect to* θ *on the whole* Θ. *Assume also that the partial derivatives* $\partial f_\theta(x)/\partial\theta_i$, $i = 1, \ldots, k$, *are uniformly bounded in* θ *and* x. *Suppose that* $\|\hat{\theta} - \theta\|_1 := \sum_{i=1}^k |\hat{\theta}_i - \theta_i| \to 0$ *a.s. as* $n \to \infty$. *Then*

a) $c_n \to c$ *a.s. as* $n \to \infty$.

b) $|P_n - \alpha| = O(\|\hat{\theta} - \theta\|_1)$ *a.s.*

c) If $(f1)$ *holds then*

$$|c_n - c| = \begin{cases} O(\|\hat{\theta} - \theta\|_1^{1/\gamma}) & a.s. \quad \text{if } \gamma > 1 \\ O(\|\hat{\theta} - \theta\|_1) & a.s. \quad \text{if } \gamma \le 1 \end{cases} \tag{5}$$

and

$$d_\theta(\{f_{\hat{\theta}} \ge c_n\}, \{f_\theta \ge c\}) = \begin{cases} O(\|\hat{\theta} - \theta\|_1) & a.s. \quad \text{if } \gamma > 1 \\ O(\|\hat{\theta} - \theta\|_1^\gamma) & a.s. \quad \text{if } \gamma \le 1. \end{cases} \tag{6}$$

Proof:

a) By the mean value theorem there exists a constant $C > 0$ such that

$$|f_{\hat{\theta}}(x) - f_\theta(x)| \le C\|\hat{\theta} - \theta\|_1 \tag{7}$$

uniformly in x. If there exists a set A with positive probability on which convergence of c_n to c does not hold, then there exists a $\delta > 0$ and a subsequence c_{n_j} such that $|c_{n_j} - c| > \delta$ for all j.

If $c_{n_j} > c + \delta$ then

$$1 - \alpha = \int_{\{f_{\hat{\theta}} \ge c_{n_j}\}} f_{\hat{\theta}} \le \int_{\{f_{\hat{\theta}} \ge c+\delta\}} f_{\hat{\theta}} \le$$

$$\leq \int_{\{f_\theta \geq c+\delta - C\|\hat\theta-\theta\|_1\}} f_{\hat\theta}$$
$$\leq \int_{\{f_\theta \geq c+\delta - C\|\hat\theta-\theta\|_1\}} f_\theta + C\|\hat\theta-\theta\|_1 \mathrm{Leb}\{f_\theta \geq c+\delta - C\|\hat\theta-\theta\|_1\}.$$

We can take n_j sufficiently large so that $C\|\hat\theta-\theta\|_1 < \delta/2$. Then on A

$$1-\alpha \leq \int_{\{f_\theta \geq c+\delta/2\}} f_\theta + C\|\hat\theta-\theta\|_1 \mathrm{Leb}\{f_\theta \geq c+\delta/2\}. \quad (8)$$

This entails a contradiction with the definition of c. Indeed, letting $n \to \infty$ we would conclude that the first term in the right-hand side of (8) would be larger than or equal to $1-\alpha$ which contradicts the fact that, since $f_\theta(x)$ is a continuous function in x for every θ, this term is strictly smaller than $1-\alpha$. If we had assumed that $c_{n_j} < c-\delta$ we would have reached a similar contradiction.

b) Observe that by (7)

$$|P_n - \alpha| = |\int_{\{f_{\hat\theta} \geq c_n\}} f_\theta - \int_{\{f_{\hat\theta} \geq c_n\}} f_{\hat\theta}| \leq \int_{\{f_{\hat\theta} \geq c_n\}} |f_{\hat\theta} - f_\theta| \leq C\|\hat\theta-\theta\|_1 \mathrm{Leb}\{f_{\hat\theta} \geq c_n\}.$$

Then conclusion *b)* follows since, for n sufficiently large, $\mathrm{Leb}\{f_{\hat\theta} \geq c_n\}$ is bounded almost surely. To see this notice that, with probability one, for n large

$$\mathrm{Leb}\{f_{\hat\theta} \geq c_n\} \leq \frac{2}{c} c_n \mathrm{Leb}\{f_{\hat\theta} \geq c_n\} \leq \frac{2}{c} \int_{\{f_{\hat\theta} \geq c_n\}} f_{\hat\theta} = 1-\alpha.$$

c) Let $M > 0$ be a sufficiently large constant. By (7) if $\gamma > 1$ then, with probability one and for n large enough,

$$\int_{\{f_{\hat\theta} \geq c+M\|\hat\theta-\theta\|_1^{1/\gamma}\}} f_{\hat\theta}$$
$$\leq \int_{\{f_\theta \geq c+M\|\hat\theta-\theta\|_1^{1/\gamma} - C\|\hat\theta-\theta\|_1\}} f_\theta +$$

$$
\begin{aligned}
&+C\mathrm{Leb}\{f_\theta \geq c + M\|\hat{\theta}-\theta\|_1^{1/\gamma} - C\|\hat{\theta}-\theta\|_1\}\|\hat{\theta}-\theta\|_1 \\
&\leq 1-\alpha - \int_{\{c\leq f_\theta \leq c+M\|\hat{\theta}-\theta\|_1^{1/\gamma}-C\|\hat{\theta}-\theta\|_1\}} f_\theta + C_1\|\hat{\theta}-\theta\|_1 \\
&\leq 1-\alpha - C_2(M\|\hat{\theta}-\theta\|_1^{1/\gamma} - C\|\hat{\theta}-\theta\|_1)^\gamma + C_1\|\hat{\theta}-\theta\|_1 \\
&\leq 1-\alpha - \|\hat{\theta}-\theta\|_1[C_2(M-1)^\gamma - C_1] < 1-\alpha,
\end{aligned}
$$

where $C_1, C_2 > 0$ are constants. This implies that $c_n < c + M\|\hat{\theta}-\theta\|_1^{1/\gamma}$. In a similar way it is easy to check that for large n

$$
\int_{\{f_{\hat{\theta}} \geq c - M\|\hat{\theta}-\theta\|_1^{1/\gamma}\}} f_{\hat{\theta}} > 1-\alpha \quad \text{a.s.},
$$

from which we obtain that $c_n > c - M\|\hat{\theta}-\theta\|_1^{1/\gamma}$.

Analogously, if $\gamma \leq 1$, with probability one we have that

$$
\int_{\{f_{\hat{\theta}} \geq c + M\|\hat{\theta}-\theta\|_1\}} f_{\hat{\theta}} \leq 1-\alpha - C_3[(M-C)\|\hat{\theta}-\theta\|_1]^\gamma + C_4\|\hat{\theta}-\theta\|_1 < 1-\alpha
$$

and

$$
\int_{\{f_{\hat{\theta}} \geq c - M\|\hat{\theta}-\theta\|_1\}} f_{\hat{\theta}} \geq 1-\alpha + C_5[(M-C)\|\hat{\theta}-\theta\|_1]^\gamma - C_6\|\hat{\theta}-\theta\|_1 > 1-\alpha,
$$

where the C_i's are positive constants. This proves (5). Regarding (6) with probability one, for n sufficiently large,

$$
\begin{aligned}
\int_{\{f_{\hat{\theta}} \geq c_n\}\Delta\{f_\theta \geq c\}} f_\theta &= \int_{\{f_{\hat{\theta}} \geq c_n\}\cap\{f_\theta < c\}} f_\theta + \int_{\{f_{\hat{\theta}} < c_n\}\cap\{f_\theta \geq c\}} f_\theta \\
&\leq \int_{\{c_n + C\|\hat{\theta}-\theta\|_1 \geq f_\theta \geq c_n - C\|\hat{\theta}-\theta\|_1\}} f_\theta \leq \int_{\{|c_n - c| + C\|\hat{\theta}-\theta\|_1 \geq |f_\theta - c|\}} f_\theta \\
&= O(\max(|c_n - c|, \|\hat{\theta}-\theta\|_1))^\gamma
\end{aligned}
$$

which completes the proof.

$\square$

Remark: Under the same conditions as in (c), it is straightforward to see that $\mathrm{Leb}(\{f_{\hat{\theta}} \geq c_n\} \Delta \{f_\theta \geq c\}) = O(\max(|c_n - c|, \|\hat{\theta} - \theta\|_1))^\gamma$ a.s. Thus the tolerance region $\{f_{\hat{\theta}} \geq c_n\}$ is also a.s. asymptotically optimal in the sense of Chatterjee and Patra (1980).

A simulation study with normal mixtures

We have carried out a simulation study in order to check the performance of the detection scheme introduced above, and to compare the differences between using a parametric and a nonparametric density estimator f_n in the tolerance region $\{f_n > c_n\}$. The original density generating the data is a mixture of Gaussian densities $f_\theta = \sum_{i=1}^k \pi_i \phi_{\mu_i, \Sigma_i}$, where $\phi_{\mu,\Sigma}$ denotes the normal density with mean μ and covariance matrix Σ, $0 \leq \pi_i \leq 1$, $i = 1, \ldots, k$, and k is a positive integer. The number k of components in the simulated mixtures was either $k = 1$, 2 or 3. We have considered two possible estimators of f_θ to plug in the tolerance region $\{f_n \geq c_n\}$. The parametric one is a mixture of normals, $f_{\hat{\theta}} = \sum_{i=1}^k \hat{\pi}_i \phi_{\hat{\mu}_i, \hat{\Sigma}_i}$, where $\hat{\pi}_i$, $\hat{\mu}_i$ and $\hat{\Sigma}_i$ denote the ML-estimators of π_i, μ_i and Σ_i respectively. To compute these estimators we used the EM algorithm (Dempster, Laird and Rubin 1977) implemented in a modification of some Matlab functions developed by Cappé (2001).

The second density estimator is of kernel type, $\hat{f}_n(x) := n^{-1} \sum_{i=1}^n K_h(x - X_i)$, where $K_h(x) := h^{-d} K(x/h)$, K is a probability density (the kernel) and $h = h_n$ is a sequence of smoothing parameters such that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. We have taken the standard Gaussian density as the kernel function K. The kernel estimator is thus a smooth density constructed as an average of rescaled normal densities placed on the observations (for a review on nonparametric density estimation see Silverman 1986 and Simonoff 1996). The degree of smoothness of the estimator is controlled by the bandwidth h, which in our study has been chosen in two different ways. First we have computed the smoothing parameter h_{PI} via a plug-in method proposed by Sheather and Jones (1991) in the univariate case $d = 1$, and studied by Wand and Jones (1994) and Duong and Hazelton (2003) in the multivariate case $d \geq 1$. The aim of the second bandwidth choice is to better control the probability of false alarm, which has full sense in this context of statistical tolerance regions. The idea is to take the bandwidth h_{CV} which locally minimizes (in a neighbourhood of h_{PI}) a cross-validation estimate of $|P_n(h) - \alpha|$ where

$$P_n(h) := \int_{\{\hat{f}_n < c_n\}} f$$

denotes the probability of false alarm. Concretely $P_n(h)$ is approximated by

$$P_{CV}(h) = \frac{1}{n} \sharp \{i = 1, \ldots, n : \hat{f}_{-i}(X_i) < c_{-i}\},$$

where $\hat{f}_{-i}$ denotes the kernel estimator, with bandwidth h and kernel K, constructed from the sample $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and c_{-i} verifies

$$\int_{\{\hat{f}_{-i} \geq c_{-i}\}} \hat{f}_{-i} = 1 - \alpha. \tag{9}$$

The window h_{CV} is defined as the median of the bandwidths h minimizing $|P_{CV}(h) - \alpha|$ over a 51-point equally spaced grid with center h_{PI} and width $1.2h_{PI}$. We have used the Matlab codes developed by Marron (1996) for the one-dimensional case. All the files used in this work are available upon request from the authors.

In the experiment for $d = 1$ (resp. $d = 2$) we have generated 300 samples of size $n = 50$ (resp. $n = 100$) from the original mixture f_θ. The constant α was fixed to 0.05. In each of the 300 repetitions we took 500 additional observations from f_θ and checked whether they belonged or not to the estimated level set $\{f_n > c_n\}$. This yielded 300 proportions of false alarms whose average value $\hat{P}_n$ was used as an approximation to the probability of false alarm P_n. Table 1 and 2 display the average values (over the 300 repetitions) of $(P_n - \alpha)/\alpha$, h_{PI} and h_{CV} for the one and two dimensional case respectively.

The other important aspect is the power, that is, the ability to properly raise an alarm. A usual way to evaluate it is the so-called average run length (ARL), the expected number of observations coming from the new (post-change) distribution recorded before raising the alarm (see e.g. Yakir 1996). In the simulations we have chosen a post-change density $f_{\theta'}$ corresponding to a shift in any of the means μ_i of the mixture or an increase in the variability or just a change in the weights w_i. In Table 1 and 2 we have marked in bold type the parameters whose value has changed from the original to the post-change distribution. In each of the 300 repetitions we have drawn 500 observations from $f_{\theta'}$ and computed the proportion of alarms. The average of these 300 proportions is the approximation to the power used to compute the ARL.

The results appearing in Table 1 correspond to the following normal mixtures f_θ (see Figure 1): standard normal N(0,1); separated bimodal 0.5 N(0,1) + 0.5 N(8,1); skewed bimodal 0.8 N(1,1) + 0.2 N(5,2) and 0.6 N(1,1) + 0.4 N(4.5,1.5); skewed unimodal 0.5 N(1,1)+ 0.5 N(3,1.5) and 0.9 N(1,1) + 0.1 N(3.5,2); kurtotic 0.70 N(0.0,1.0) + 0.30

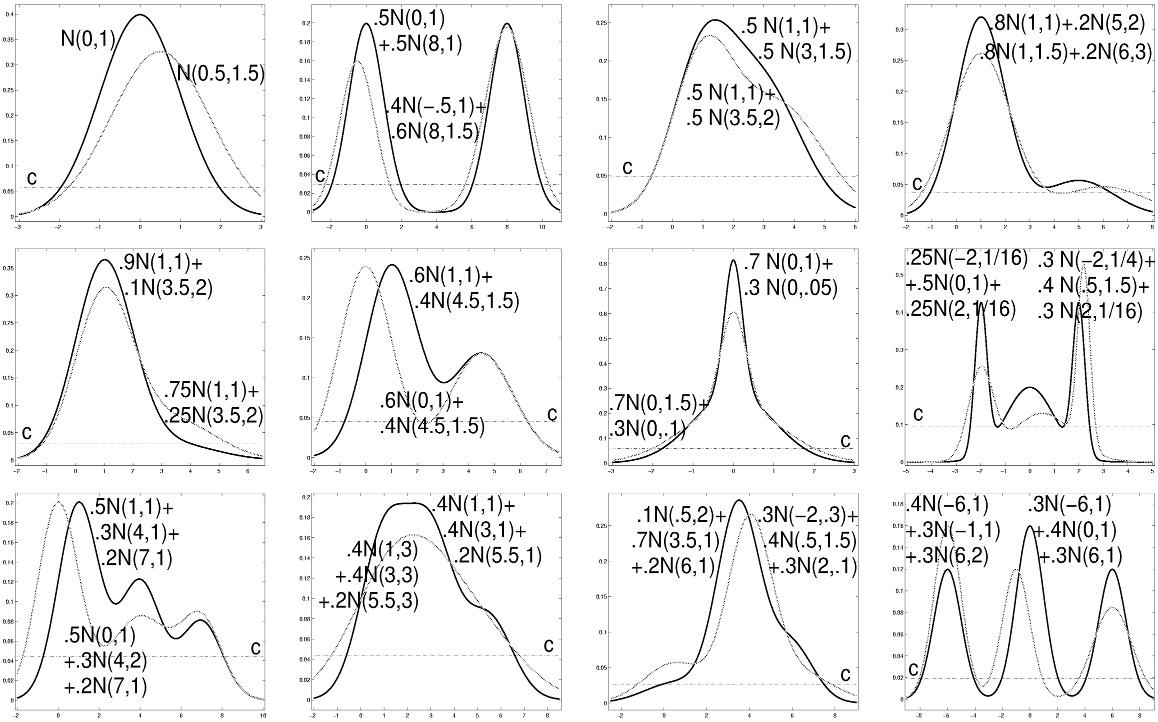


Figure 1: Pre-change (dark line) and post-change (light line) normal mixtures in the simulations for $d = 1$. The dashed line corresponds to the c level determined by (2).

N(0.0,0.05); trimodal 0.25 N(-2,1/16) + 0.5 N(0,1) + 0.25 N(2,1/16); skewed trimodal 0.5 N(1,1) + 0.3 N(4,1) + 0.2 N(7,1); skewed unimodal 0.4 N(1,1) + 0.4 N(3,1) + 0.2 (5.5,1) and 0.1 N(0.5,2) + 0.7 N(3.5,1) + 0.2 N(6,1) and separated trimodal 0.3 N(-6,1) + 0.4 N(0,1) + 0.3 N(6,1).

Our main conclusions are the following:

i) The mixture estimator is the one with smallest ARL in every case, so it raises the alarm most frequently when it is due. But at the same time it always has a higher probability of false alarm than it should, in fact, the furthest of the three from the target value $\alpha = 0.05$. In any case, we should always bear in mind that the parametric method headed off with a slight advantage since we have prespecified

f_θ	f_n	$(\hat{P}_n - \alpha)/\alpha$	h	$f_{\theta'}$	ARL
	Mixture	.1643			5.68
N(0,1)	Kernel h_{PI}	-.1451	.4485	N(**0.5**,**1.5**)	7.28
	Kernel h_{CV}	-.0173	.3929		6.63
0.5 N(0,1)	Mixture	.4805		**0.4** N(**-0.5**,1)	7.04
+ 0.5 N(8,1)	Kernel h_{PI}	-.4556	.7199	+ **0.6** N(8,**1.5**)	18.63
	Kernel h_{CV}	-.0013	.5184		10.30
0.5 N(1,1)	Mixture	.2675		0.5 N(1,1)	7.68
+ 0.5 N(3,1.5)	Kernel h_{PI}	-.3009	.6770	+ 0.5 N(**3.5**,**2**)	12.17
	Kernel h_{CV}	-.1016	.5439		10.07
0.8 N(1,1)	Mixture	.8172		0.8 N(1,**1.5**)	5.81
+ 0.2 N(5,2)	Kernel h_{PI}	.1677	.6093	+ 0.2 N(**6**,**3**)	7.74
	Kernel h_{CV}	.1092	.6741		8.01
0.9 N(1,1)	Mixture	.4085		**0.75** N(1,1)	6.76
+ 0.1 N(3.5,2)	Kernel h_{PI}	.0352	.5210	+ **0.25** N(3.5,2)	8.67
	Kernel h_{CV}	-.0096	.5470		8.95
0.6 N(1,1)	Mixture	.3644		0.6 N(**0**,1)	5.20
+ 0.4 N(4.5,1.5)	Kernel h_{PI}	-.2268	.7019	+0.4 N(4.5,1.5)	8.39
	Kernel h_{CV}	.0025	.5704		6.81
0.7 N(0,1)	Mixture	.5583		0.7 N(0,**1.5**)	6.65
+ 0.3 N(0,0.05)	Kernel h_{PI}	.3900	.2688	+ 0.3 N(0,0.05)	7.43
	Kernel h_{CV}	.2952	.3092		7.90
0.25 N(-2,1/16)	Mixture	.7127		**0.3** N(-2,**1/4**)	4.81
+ 0.5 N(0,1)	Kernel h_{PI}	-.3412	.3503	+ **0.4** N(**0.5**,**1.5**)	9.89
+ 0.25 N(2,1/16)	Kernel h_{CV}	.0749	.2657	+ **0.3** N(**2.2**,1/16)	7.14
0.5 N(1,1)	Mixture	.7924		0.5 N(**0**,1)	4.63
+ 0.3 N(4,1)	Kernel h_{PI}	-.2688	.8778	+ 0.3 N(4,**2**)	10.22
+ 0.2 N(7,1)	Kernel h_{CV}	-.0161	.6872	+ 0.2 N(7,1)	7.65
0.4 N(1,1)	Mixture	.5659		0.4 N(1,**3**)	5.03
+ 0.4 N(3,1)	Kernel h_{PI}	-.3045	.8389	+ 0.4 N(3,**3**)	8.34
+ 0.2 N(5.5,1)	Kernel h_{CV}	-.0565	.6410	+ 0.2 N(5.5,**3**)	6.89
0.1 N(0.5,2)	Mixture	.7681		**0.2** N(0.5,2)	6.50
+ 0.7 N(3.5,1)	Kernel h_{PI}	.1547	.6908	+ **0.6** N(**4**,1)	8.68
+ 0.2 N(6,1)	Kernel h_{CV}	.1084	.7340	+ 0.2 N(6,**3**)	9.00
0.3 N(-6,1)	Mixture	.7783		**0.4** N(-6,1)	3.55
+ 0.4 N(0,1)	Kernel h_{PI}	-.4002	.8057	+ **0.3** N(**-1**,1)	7.83
+ 0.3 N(6,1)	Kernel h_{CV}	-.0069	.6586	+ 0.3 N(6,**2**)	6.50

Table 1: Simulation outputs for the one-dimensional case.

f_θ	f_n	$(\hat{P}_n-\alpha)/\alpha$	h	$f_{\theta'}$	ARL
Standard normal $\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\&1\end{bmatrix}\right)$	Mixture K h_{PI} K h_{CV}	.1669 -.0129 .1003	 .4488 .4318	$\mathrm{N}\left(\begin{bmatrix}\mathbf{.5}\\0\end{bmatrix},\begin{bmatrix}1&0\\&\mathbf{1.5}\end{bmatrix}\right)$	7.24 8.57 7.83
Separated bimodal $.5\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\&1\end{bmatrix}\right)$ $+.5\mathrm{N}\left(\begin{bmatrix}4\\0\end{bmatrix},\begin{bmatrix}1&0\\&1\end{bmatrix}\right)$	Mixture K h_{PI} K h_{CV}	.2563 -.4881 .0185	 .6682 .5094	$.5\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\mathbf{2}&0\\&\mathbf{2}\end{bmatrix}\right)$ $+.5\mathrm{N}\left(\begin{bmatrix}\mathbf{4.5}\\0\end{bmatrix},\begin{bmatrix}1&0\\&1\end{bmatrix}\right)$	5.50 10.30 6.49
Skewed bimodal $.6\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&.5\\&1.5\end{bmatrix}\right)$ $+.4\mathrm{N}\left(\begin{bmatrix}3\\0\end{bmatrix},\begin{bmatrix}1.5&-.5\\&1.5\end{bmatrix}\right)$	Mixture K h_{PI} K h_{CV}	.4013 -.2765 .0757	 .6525 .5519	$\mathbf{.5}\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\mathbf{1.5}&\mathbf{.3}\\&1.5\end{bmatrix}\right)$ $+\mathbf{.5}\mathrm{N}\left(\begin{bmatrix}3\\0\end{bmatrix},\begin{bmatrix}1.5&-.5\\&\mathbf{2}\end{bmatrix}\right)$	7.64 14.46 10.12
Skewed unimodal $.7\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\&1\end{bmatrix}\right)$ $+.3\mathrm{N}\left(\begin{bmatrix}2.5\\0\end{bmatrix},\begin{bmatrix}1.5&.5\\&1.5\end{bmatrix}\right)$	Mixture K h_{PI} K h_{CV}	.4364 -.0960 .0296	 .5733 .5461	$\mathbf{.6}\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\&\mathbf{2}\end{bmatrix}\right)$ $+\mathbf{.4}\mathrm{N}\left(\begin{bmatrix}\mathbf{3.5}\\0\end{bmatrix},\begin{bmatrix}1.5&.5\\&1.5\end{bmatrix}\right)$	4.06 5.93 5.37
Separated trimodal $.4\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1.5&0\\&1\end{bmatrix}\right)$ $+.3\mathrm{N}\left(\begin{bmatrix}5\\0\end{bmatrix},\begin{bmatrix}1&0\\&1.5\end{bmatrix}\right)$ $+.3\mathrm{N}\left(\begin{bmatrix}2.5\\3\end{bmatrix},\begin{bmatrix}1&.2\\&1\end{bmatrix}\right)$	Mixture K h_{PI} K h_{CV}	.5981 -.5573 .0391	 .8162 .6004	$.4\mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\mathbf{2}&\mathbf{.5}\\&\mathbf{1.5}\end{bmatrix}\right)$ $+.3\mathrm{N}\left(\begin{bmatrix}\mathbf{4.5}\\0\end{bmatrix},\begin{bmatrix}1&0\\&1.5\end{bmatrix}\right)$ $+.3\mathrm{N}\left(\begin{bmatrix}\mathbf{3}\\3\end{bmatrix},\begin{bmatrix}1&.2\\&1\end{bmatrix}\right)$	9.26 29.77 13.81

Table 2: Simulation outputs for the two-dimensional case.

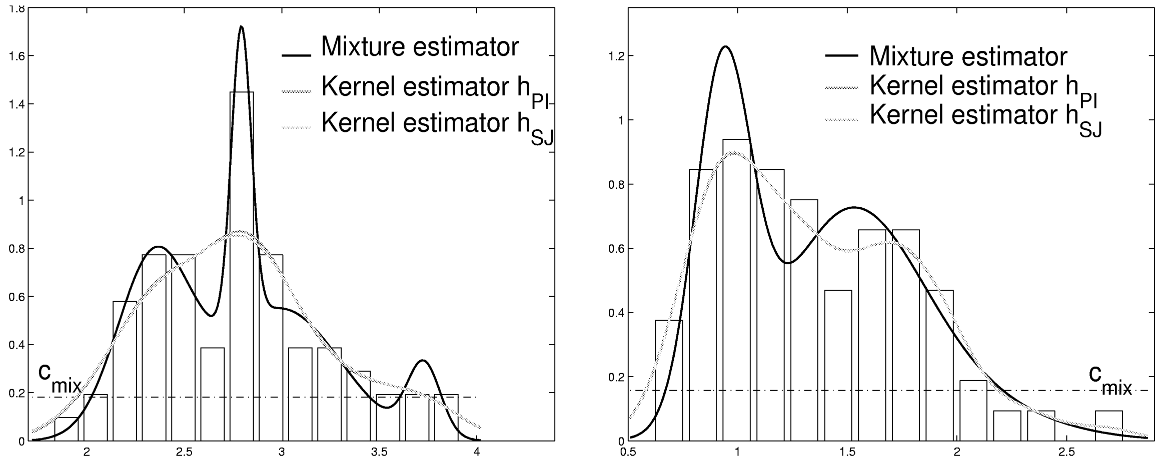
the correct number of components k in the estimated mixture. Obviously, this will not always be possible when dealing with real data problems (see Leroux 1992 and references therein for effects of overparametrizing an estimated mixture).

ii) When the components in the mixture are not clearly distinguishable the computational difficulties associated with the estimation of a large number of parameters are more apparent. The nonparametric estimator automatically adapts to these situations without any remarkable loss in its performance.

iii) The cross-validation choice h_{CV} for the bandwidth is the one providing a probability of false alarm nearest to $\alpha = 0.05$. Observe that even though this probability is generally lower for the window h_{PI}, the power is also lower for this choice. The bandwidth h_{CV} reaches an adequate "balance" between the probability of the two types of errors: false alarm and undetected change.

A case study: amino-acids in citrus juice

Here we consider a case study analyzed in Fuchs and Kenett (1998). The data set consists of the concentrations in miug per standard volume of eleven amino-acids (such as lysine, alanine, aspartic acid, ...) in 69 extractions from a pure juice. The quality of further samples of fruit juice is checked by comparing their constituents with those of this base sample. Fuchs and Kenett (1998) review the construction of multivariate tolerance regions under the assumption of normality as a means to detect a possible adulteration in the pure juice. However some variables have a clear departure from



(a) (b)

Figure 2: Concentration of aspartic acid (a) and alanine (b) in juice data from Fuchs and Kenett (1998).

normality and, due to the presence of more than one mode, they cannot be transformed to resemble a Gaussian distribution (see Figure 2). A possible reason why this happens is that the values of the attributes differ depending on varieties of the same fruit or environmental and climatic factors.

As before we are interested in checking the performance of a mixture of normal densities and a kernel density estimator f_n in the detection procedure provided by tolerance regions of the type $\{f_n \geq c_n\}$. We have used the Bayesian information criterion (see, e.g., McLachlan and Peel 2000) to decide on the smallest number k of components compatible with the data. The evaluation of the probability of false alarm and the power was carried out via a leave-one-out procedure: each observation X_i, $i = 1, \ldots, 69$, was successively dropped from the sample and a density estimator $\hat{f}_{-i}$ was constructed from sample $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{69}$. The estimated probability of false alarm is given by

$$\hat{p}_{FA} = \frac{1}{69}\sum_{i=1}^{69} 1_{\{X_i \in \{f_{-i} < c_{-i}\}\}},$$

where c_{-i} is determined in expression (9). The process of adulteration was simulated by multiplying each X_i by a reduction factor of 0.9 or 0.75 or by subtracting a constant (0.5 or 1). If we denote the resulting observation by Y_i the probability of detecting the change is estimated by

$$\hat{p}_{pow} = \frac{1}{69}\sum_{i=1}^{69} 1_{\{Y_i \in \{f_{-i} < c_{-i}\}\}}.$$

	Kernel h_{PI}	Kernel h_{CV}	Mixture	
$\hat{c}$	.1601	.1573	.1809	
$\hat{p}_{FA}$	.0435	.0580	.1304	
ARL	1.76	1.88	0.30	$0.75X$
	16.24	16.24	0.60	$0.9X$
	0.50	0.50	0.19	$X-1$
	2.83	3.06	0.33	$X-0.5$
h	.1830	.1924		

(a)

	Kernel h_{PI}	Kernel h_{CV}	Mixture	
$\hat{c}$	.0564	.0662	.0512	
$\hat{p}_{FA}$	.0145	.0435	.0725	
ARL	2.83	2.14	2.29	$0.75X$
	21.99	16.24	10.49	$0.9X$
	1.65	1.16	0.60	$X-0.5$
h	.2012	.1715		

(b)

Table 3: Tolerance regions detection procedure applied to the concentration of (a) aspartic acid , (b) aspartic acid and alanine, in juice data from Fuchs and Kenett (1998)

The results are displayed in Table 3: in (a) the data X_i were the concentrations of aspartic acid (univariate); in (b) the X_i's were the concentrations of aspartic acid and alanine (bivariate). In general the parametric estimator outperforms the nonparametric one when a real change has taken place in the distribution of the process. However the kernel estimator manages to keep the probability of false alarm around the maximum desired value, $\alpha = 0.05$, while this probability is excessively high in the case of the estimated mixture.

References

Aitchison, J. and Dunsmore, I. R. (1975), *Statistical Prediction Analysis.* Cambridge University Press.

Baíllo, A., Cuevas, A. and Justel, A. (2000), Set estimation and nonparametric detection. *Canad. J. Statist.*, 28, 765-782.

Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001), Convergence rates in nonparametric estimation of level sets. *Statist. Prob. Letters*, 53, 27-35.

Baíllo, A. (2003), Total error in a plug-in estimator of level sets. *Statist. Prob. Letters*, 65, 411-417.

Bucchianico A. Di, Einmahl, J. H. J. and Mushkudiani, N. (2001), Smallest nonparametric tolerance regions. *Ann. Statist.*, 29, 5, 1320-1343.

Chatterjee, S. K. and Patra, N. K. (1980), Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull.*, 29, 73-93.

Crosier, R. B. (1988), Multivariate generalizations of cumulative sum quality control schemes. *Technometrics*, 30, 3, 291-303.

Csörgö, M. and Horváth, L. (1997), *Limit Theorems in Change-Point Analysis.* Wiley, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39, 1, 1-38.

DeVor, R. E., Chang, T.-H. and Sutherland, J. W. (1992), *Statistical Quality Design and Control.* McMillan, New York.

Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: the L^1 View.* Wiley, New York.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38, 480–488.

Duong, T. and Hazelton, M. L. (2003), Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametric Statist.*, 15, 17-30.

Fuchs, C. and Kenett, R. S. (1987), Multivariate tolerance regions and F-tests. *J. Quality Tech.*, 19, 122-131.

Fuchs, C. and Kenett, R. S. (1988), Appraisal of ceramic substrates in multivariate tolerance regions. *The Statistician*, 37, 401-411.

Fuchs, C. and Kenett, R. S. (1998), *Multivariate Quality Control. Theory and Applications.* Marcel Dekker, New York.

Gordon, L. and Pollak, M. (1995), A robust surveillance scheme for stochastically ordered alternatives. *Ann. Statist.*, 23, 4, 1350-1375.

Guttman, I. (1970), *Statistical Tolerance Regions: Classical and Bayesian.* Griffin, London.

Hawkins, D. M. and Olwell, D. H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement.* Springer-Verlag, New York.

Healy, J. D. (1987), A note on multivariate CUSUM procedures. *Technometrics*, 29, 409-412.

Hotelling, H. (1947), Multivariate quality control illustrated by air testing of sample bombsights. *Selected Techniques of Statistical Analysis.* C. Eisenhart et al. editors. Mc Graw-Hill, New York.

Leroux, B. G. (1992), Consistent estimation of a mixing distribution. *Ann. Statist.*, 20, 1350-1360.

Liu, R. Y. and Singh, K. (1993), A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* 88, 421, 252-260.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models.* Wiley, New York.

Mushkudiani, N. (2000), *Statistical Applications of Generalized Quantiles: Nonparametric Tolerance Regions and P-P Plots.* Eindhoven University of Technology, Eindhoven.

Page, E. S. (1954), Continuous inspection schemes. *Biometrika*, 41, 100-115.

Polansky, A. M. (2001), A smooth nonparametric approach to multivariate process capability. *Technometrics*, 43, 2, 199-211.

Pollak, M. and Siegmund, D. (1985). A diffusion process and its application to detecting a change in the drift of Brownian motion. *Biometrika*, 72, 267-280.

Roberts, S. W. (1966), A comparison of some control chart procedures. *Technometrics*, 8, 411-430.

Sheather, S. J. and Jones, M. C. (1991), A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 53, 683-690.

Shewhart, W. A. (1931), *The Economic Control of the Quality of Manufactured Product.* Macmillan, New York.

Shiryaev, A. N. (1963), On optimum methods in quickest detection problems. *Theory Probab. Appl.*, 8, 22-46.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics.* Springer-Verlag, New York.

Tsybakov, A. B. (1997), On nonparametric estimation of density level sets. *Ann. Statist.* 25, 948-969.

Wand, M. P. and Jones, M. C. (1994), Multivariate plug-in bandwidth selection. *Comput. Statist.*, 9, 97-116.

Wetherill, G. B. and Brown, D. W. (1994), *Statistical Process Control.* Chapman and Hall, London.

Yakir, B. (1996), A lower bound on the ARL to detection of a change with a probability constraint on false alarm. *Ann. Statist.*, 24, 431-435.